PMI-cool at SemEval-2016 Task 3: Experiments with PMI and Goodness Polarity Lexicons for Community Question Answering

Daniel Balchev, Yasen Kiprov, Ivan Koychev
Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
Sofia, Bulgaria
{dblachev,yasen.kiprov}@gmail.com
koychev@fmi.uni-sofia.bg

Preslav Nakov
Qatar Computing Research Institute
HBKU, P.O. box 5825
Doha, Qatar
pnakov@qf.org.qa

Abstract

We describe our submission to SemEval-2016 Task 3 on Community Question Answering. We participated in subtask A, which asks to rerank the comments from the thread for a given forum question from good to bad. Our approach focuses on the generation and use of goodness polarity lexicons, similarly to the sentiment polarity lexicons, which are very popular in sentiment analysis. In particular, we use a combination of bootstrapping and pointwise mutual information to estimate the strength of association between a word (from a large unannotated set of question-answer threads) and the class of good/bad comments. We then use various features based on these lexicons to train a regression model, whose predictions we use to induce the final comment ranking. While our system was not very strong as it lacked important features, our lexicons contributed to the strong performance of another top-performing system.

1 Introduction

Online forums have been gaining a lot of popularity in recent years. In these forums, one can ask a question, and based on the wisdom of the crowd, expect to get some good answers. In practice, unless there is strong moderation, most such forums get populated with bad answers, which can be annoying for users as it takes time to read through all answers in a long thread. As the importance of the problem was recognized in the research community, this gave rise to two shared tasks on Community Question Answering at SemEval-2015 (Nakov et al., 2015) and SemEval-2016 (Nakov et al., 2016).

Here we describe the PMI-cool system, which we developed to participate in SemEval-2016 Task 3, subtask A, which asks to rerank the answers in a question-answer thread, ordering them from good to bad (Nakov et al., 2016). As the name of our system suggests, our approach is heavily based on pointwise mutual information (PMI), which we use to estimate the association strength between a word and a class, e.g., the class of good or the class of bad comments. Based on this association strength, we perform bootstrapping in a large unannotated set of question-answer threads to generate specialized *goodness polarity* lexicons. We then use various features based on these lexicons to train a regression model, whose predictions we use to induce the final comment ranking.

While our PMI-cool system did not perform very well at the competition as it lacked important features and as we found a bug in our submission, our goodness polarity lexicons proved useful and contributed to the strong performance of another top-performing system at SemEval-2016 Task 3: SUper_team (Mihaylova et al., 2016).

2 Method

Our solution can be separated into two phases: (*i*) feature extraction, and (*ii*) machine learning. The feature extraction phase consists of extracting various PMI-based and other features, which we describe in the following sections. In the second phase, we apply a support vector machine (SVM) regression model (Drucker et al., 1997), taking the features as an input and returning the similarity score for each question-answer pair as an output.

Proceedings of SemEval-2016, pages 769–775,
San Diego, California, June 16-17, 2016.

At test time, we generate regression scores for each answer in a question-answer thread and we rerank the answers accordingly. Before exploring our features, we will first introduce PMI and how we use it to generate goodness polarity lexicons.

3 Pointwise Mutual Information and Strength of Association

The pointwise mutual information (PMI) is a notion from the theory of information: given two random variables A and B, the mutual information of A and B is the "amount of information" (in units such as bits) obtained about the random variable A, through the random variable B (Church and Hanks, 1990).

Let a and b be two values from the sample space of A and B, respectively. The *pointwise* mutual information between a and b is defined as follows:

$$pmi(a;b) = \log \frac{P(A=a, B=b)}{P(A=a) \cdot P(B=b)} \quad (1)$$

$$= \log \frac{P(A=a|B=b)}{P(A=a)} \quad (2)$$

$pmi(a;b)$ takes values between $-\infty$, which is when $P(A = a, B = b) = 0$, and $\min\{-\log P(A=a), -\log P(B=b)\}$, when $P(A=a|B=b) = P(B=b|A=a) = 1$.

The mutual information between A and B is the expected value of $pmi(a;b)$:

$$MI(A,B) = \sum_{a \in A} \sum_{b \in B} pmi(a;b) \quad (3)$$

PMI is central to a popular approach for bootstrapping sentiment lexicons proposed by Turney (2002). The idea is to start with a small set of seed positive (e.g., *excellent*) and negative words (*bad*), and then to use these words to induce sentiment polarity orientation for new words in a large unannotated set of texts (in his case, product reviews). The idea is that words that co-occur in the same text with positive seed words are likely to be positive, while those that tend to co-occur with negative words are likely to be negative. To quantify this intuition, Turney defines the notion of semantic orientation (SO) for a term w as follows:

$SO(w) = pmi(w, pos) - pmi(w, neg)$

where pos and neg stand for any positive and negative seed word, respectively.

The idea was later used by other researchers, e.g., Mohammad et al. (2013) built several lexicons based on PMI between words and tweet categories. Here the categories (positive and negative) were defined by a seed set of emotional hashtags, e.g., #happy, #sad, #angry, etc. or by simple positive and negative smileys, e.g., `;)`, `:)`, `;(`, `:(`. In this case, the resulting lexicons included not only words, but also bigrams and discontinuous pairs of words.

Another related work is that of Severyn and Moschitti (2015), who proposed an approach to lexicon induction, which, instead of using PMI for SO, assigns positive/negative labels to the unlabeled tweets (based on the seeds), and then trains an SVM classifier on them, using word n-grams as features. These n-grams are then used as lexicon entries with the learned classifier weights as polarity scores. While this is an interesting approach, in our experiments below, we will stick to PMI as a more established method to estimate SO.

Finally, there is a related task at SemEval-2016 on predicting the out-of-context sentiment intensity of phrases (Kiritchenko et al., 2016), but there the focus is on multiword phrases.

4 Building Goodness Polarity Lexicons

We use SO to build goodness polarity lexicons for good/bad comments in the forum. Instead of using positive and negative sentiment words as seeds, we start with seed words that are associated with good or bad comments. Unlike the work above, we do not do pure bootstrapping, but rather we use a semi-supervised approach, which works in two steps.

Step 1: In order to come up with a list of words that signal a good/bad comment (which is not as easy as it is to come up with such words manually), we look for words that are strongly associated with good vs. bad comments in the annotated training dataset, using SO. We then select the top 5% of the words with the most extreme positive/negative values of SO, which corresponds to the most extreme good/bad comment words.

Step 2: We then apply the SO again, but this time using the seed words that we selected in Step 1, in order to build the final large-scale goodness polarity lexicon, as in the above-described work.

5 Features

We used a variety of features based on the textual content of the question and of the answer and on metadata about the question and about the question-answer pair.

5.1 Metadata features

All the metadata features we used are included with their SO with the good/bad class. We used the following features:

- *SameAuthor.* This feature checks whether the target answer is given by the same user who asked the question. The assumption here is that the author of the question is unlikely to provide a good answer to his/her own question. We do not use this boolean feature directly, but we use the SO between it and the good/bad classes.

- *AnswerNumber.* This is the rank of the answer (e.g., first, second, third, ..., tenth). The assumption is that most discussions tend to degenerate and to lose focus over time. This is also visible in the baseline that ranks the answers based on their chronologicak rank, which performs better than random (Nakov et al., 2016). The feature value is the SO of the answer rank and of the good/bad classes.

- *AnswerAuthor.* This is the ID of the person who gave the answer. The idea is that some users might tend to give mostly good/bad answers. Thus, the SO between the author ID and the good/bad classes is useful for user modeling.

5.2 Word PMI

The main feature of our PMI-cool system is based on the lexicon constructed by computing the SO of each word, used in all the answers in the training corpus. Using this technique, we identify words commonly used in good versus bad answers in general, regardless of the question; we used words without stemming as stemming lowered the performance. For instance, bad answers often contain variants of thanks statements, insults, words generally used in off-topic comments and interjections, etc. Table 1 shows some of the top words that are most strongly associated with bad answers in terms of SO.

5.3 Sentiment Lexicon

Another resource we used is the Sentiment140 lexicon, which was constructed by Mohammad et al. (2013) using SO for word weighting, as we mentioned above. Our assumption here is that good/bad sentiment expressed in the answer suggests good/bad answers, as previously suggested in (Nicosia et al., 2015). The feature we calculate is the sum of the sentiment scores of the sentiment-bearing words in the answer.

5.4 Bootstrapped PMI

As explained above, we used bootstrapping to induce larger goodness polarity lexicons from the large unnanotated corpus provided for the task. For this purpose, we first used PMI to build a lexicon from the labeled training data, removing rarely mentioned words and taking the top and the bottom 5% of the rest, based on the SO score. Then, we used PMI again, using these words as good/bad seeds to generate the large lexicon. Unfortunately, due to implementation issues before the submission deadline, this feature was not included in the submitted system's feature set.

6 Ineffective Features

We further used some features that turned out to be ineffective. Still, we describe them here as we believe this might be useful for other researchers.

6.1 Personality Trait Features

We used the lexicons of Schwartz et al. (2013), which were designed to measure a user's big-five personality traits: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. These lexicons were also computed using PMI and SO metrics between words in a large set of user-generated content on social media: 75,000 Facebook user profiles with personality values and 700 million words, phrases, and topic instances. There are two lexicons for each trait, with the 100 most and the 100 least characteristic words for the target trait.

We calculated a personality score for each of the five traits by summing the scores of all matching words in all answers written by the author of the answer, thus modeling his/her personality profile. However, this feature, did not yield improvements.

Word	SO
thanx	-2.1962458411
wk	-2.1638105653
tnx	-2.0627144485
thanks	-2.0458352035
thx	-1.965984822
thank	-1.9291830558
lols	-1.8324534294
colt	-1.8324534294
khanan	-1.7960857852
md	-1.7090744082
richard	-1.6989220368
khattak	-1.6783027496
huh	-1.6783027496
appreciate	-1.6643165076
avatar	-1.5798626767
joking	-1.5798626767
hahaha	-1.5798626767
tinker	-1.5798626767
gun	-1.5447713569
dracula	-1.5157838201
lp	-1.4959811928
idiot	-1.4234104999
bach	-1.4234104999
weird	-1.4141511745
valuable	-1.3906206771
illusions	-1.3906206771
ah	-1.3906206771
wonder	-1.375353205
silly	-1.3418305129
wow	-1.3334622633
fs	-1.3334622633

Table 1: Words with the smallest SO.

6.2 Topic Features

The text features used in PMI-cool are based on words only. We also tried to build a Latent Dirichlet Allocation (LDA) model (Blei et al., 2003) on the question and on the answers and to add the resulting topic distributions as features. However, this did not help on the development dataset, and thus we did not use it in our final submission.

7 Experiments and Evaluation

For training the prediction model for good versus bad answers, we used an SVM with a linear kernel as implemented in LibLinear (Fan et al., 2008). We treated each answer as a separate instance with all the above features, merging the PotentiallyUseful and the Bad labels under the *bad* class, and we ranked the answers based on the SVM score.

Features	MAP
MAX possible score	0.865
All features	0.638
All − metadata	0.620
All − personality	0.638
All − sentiment	0.636
All − word PMI	0.572
Baseline (chronological)	0.595
Baseline (random)	0.535

Table 2: Ablation results on the development dataset.

Here is a list of the useful features we experimented with:

- *SO SameAuthor*;
- *SO AnswerNumber*;
- *SO AnswerAuthor*;
- sum of $SO(w)$ for the answer words;
- sum of the sentiment for the answer words;
- one feature for each personality trait: sum of the scores of the lexicon words for that trait in all posts by the answer author;
- number of words with positive $BootstrapedSO$ in the answer;
- number of words with negative $BootstrapedSO$ in the answer;
- fraction of words with positive $BootstrapedSO$ in the answer;
- fraction of words with negative $BootstrapedSO$ in the answer;
- sum of the positive $BootstrapedSO$ scores for the answer words;
- sum of the negative $BootstrapedSO$ scores for the answer words;
- maximum value of $BootstrapedSO$ for a word in the answer;
- minimum value of $BootstrapedSO$ for a word in the answer;
- sum of $BootstrapedSO$ scores for all answer words.

The MAP scores resulting from our experiments on the development dataset are shown in Table 2. The first row shows the maximum possible score: it is lower than 1, as 33 of the 244 dev threads had no good answers. Next, we show the MAP score when all features are enabled; we can see that it outperforms both the chronological and the random baseline, by 4 and 10 MAP points absolute, respectively. The following four rows show results with some class of features disabled. We can see that the personality features had virtually no impact on the results, sentiment had a minimal impact (0.2 MAP points), metadata had a real impact (1.8 MAP points), while the word PMI features had the largest impact (6.6 MAP points).

8 Post-Submission Analysis

After the competition ended, we fixed a bug in the bootstrapped lexicon construction, which resulted in sizable improvements. We further replaced the uniting SVR with SVC, and we excluded the PotentiallyUseful comments from training. These changes collectively yielded a boost in MAP to 74.67 on the test dataset. As Table 3 shows, this is 6 MAP points absolute higher than the score for the system we submitted to the competition. It is also only 4.5 MAP points behind the best, and only 3 points behind the second-best team.

9 Conclusion and Future Work

We have described our PMI-cool system for SemEval-2016, Task 3 on Community Question Answering, subtask A, which asks to rerank the comments from the thread for a given forum question from good to bad. Our approach relied on using SO scores based on PMI to construct various features, the most important of which were our goodness polarity lexicons, which are based on an idea we borrowed from sentiment analysis. In particular, we used a combination of bootstrapping and pointwise mutual information to estimate the strength of association between a word (from a large unannotated set of question-answer threads) and the class of good/bad comments. We then used various features based on these lexicons to train a regression model, whose predictions we used to induce the final comment ranking.

While our PMI-cool system did not perform very well at the competition as it lacked important features and as we had a bug at submission time, our goodness polarity lexicons proved useful and contributed to the strong performance of another top-performing system at SemEval-2016 Task 3: SUper_team (Mihaylova et al., 2016).

In future work, we plan to strengthen our system with more features. In particular, we would like to incorporate rich knowledge sources, e.g., semantic similarity features based on fine-tuned word embeddings and topics similarities as in the SemanticZ system (Mihaylov and Nakov, 2016b). There are also plenty of interesting features to borrow from the SUper_Team system (Mihaylova et al., 2016), including veracity, text complexity, and troll user features as inspired by (Mihaylov et al., 2015a; Mihaylov et al., 2015b; Mihaylov and Nakov, 2016a). It would be interesting to combine these in a deep learning architecture, e.g., as in the MTE-NN system (Guzmán et al., 2016a; Guzmán et al., 2016b), which borrowed an entire neural network framework and achitecture from previous work on machine translation evaluation (Guzmán et al., 2015).

We further plan to use information from entire threads to make better predictions, as using thread-level information for answer classification has already been shown useful for SemEval-2015 Task 3, subtask A, e.g., by using features modeling the thread structure and dialogue (Nicosia et al., 2015; Barrón-Cedeño et al., 2015), or by applying thread-level inference using the predictions of local classifiers (Joty et al., 2015; Joty et al., 2016). How to use such models efficiently in the ranking setup of 2016 is an interesting research question.

Acknowledgments.

This research was performed by Daniel Balchev, a student in Computer Science in the Sofia University "St Kliment Ohridski", as part of his MSc thesis.

This research is also part of the Interactive sYstems for Answer Search (Iyas) project, which is developed by the Arabic Language Technologies (ALT) group at the Qatar Computing Research Institute (QCRI), HBKU, part of Qatar Foundation in collaboration with MIT-CSAIL.

	Submission	MAP	AvgRec	MRR	P	R	F1	Acc
1	SemEval 1st	79.19_1	88.82_1	86.42_1	76.96_1	55.30_8	64.36_5	75.11_2
2	SemEval 2nd	77.66_2	88.05_3	84.93_4	75.56_2	58.84_6	66.16_2	75.54_1
3	SemEval 3rd	77.58_3	88.14_2	85.21_2	74.13_4	53.05_{10}	61.84_8	73.39_5
...	...	...	...	...	...	...	...	...
	PMI-cool-improved	***74.67***	***85.28***	***83.54***	***76.43***	***33.18***	***46.27***	***68.69***
...	...	...	...	...	...	...	...	...
10	**PMI-cool-primary**	$\mathbf{68.79}_{10}$	$\mathbf{79.94}_{10}$	$\mathbf{80.00}_9$	$\mathbf{47.81}_{12}$	$\mathbf{70.58}_2$	$\mathbf{57.00}_9$	$\mathbf{56.73}_{12}$
...	...	...	...	...	...	...	...	...
12	SemEval 12th	62.24_{12}	75.41_{12}	70.58_{12}	50.28_{11}	53.50_9	51.84_{10}	59.60_{11}
	Baseline 1 (chronological)	**59.53**	**72.60**	**67.83**	—	—	—	—
	Baseline 2 (random)	52.80	66.52	58.71	40.56	74.57	52.55	45.26
	Baseline 3 (all 'true')	—	—	—	40.64	100.00	**57.80**	40.64
	Baseline 4 (all 'false')	—	—	—	—	—	**—**	**59.36**

Table 3: Comparison to the official results on SemEval-2016 Task 3, subtask A. The first column shows the rank of the primary runs with respect to the official MAP score. The second column contains the team's name and its submission type. The following columns show the results for the primary, and then for other, unofficial evaluation measures. The subindices show the rank of the primary runs with respect to the evaluation measure in the respective column.

References

Alberto Barrón-Cedeño, Simone Filice, Giovanni Da San Martino, Shafiq Joty, Lluís Màrquez, Preslav Nakov, and Alessandro Moschitti. 2015. Thread-level information for comment classification in community question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, ACL-IJCNLP '15, pages 687–693, Beijing, China.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, March.

Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alex J. Smola, and Vladimir Vapnik. 1997. Support vector regression machines. In M. I. Jordan and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 155–161. MIT Press.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2015. Pairwise neural machine translation evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, ACL-IJCNLP '15, pages 805–814, Beijing, China.

Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2016a. Machine translation evaluation meets community question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL '16, Berlin, Germany.

Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2016b. MTE-NN at SemEval-2016 Task 3: Can machine translation evaluation help community question answering? In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, USA.

Shafiq Joty, Alberto Barrón-Cedeño, Giovanni Da San Martino, Simone Filice, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2015. Global thread-level inference for comment classification in community question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, pages 573–578, Lisbon, Portugal.

Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2016. Joint learning with global inference for comment classification in community question answering. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '16, San Diego, California, USA.

Svetlana Kiritchenko, Saif M Mohammad, and Mohammad Salameh. 2016. SemEval-2016 task 7: Determining sentiment intensity of English and Arabic

phrases. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, USA.

Todor Mihaylov and Preslav Nakov. 2016a. Hunting for troll comments in news community forums. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL '16, Berlin, Germany.

Todor Mihaylov and Preslav Nakov. 2016b. SemanticZ at SemEval-2016 Task 3: Ranking relevant answers in community question answering using semantic similarity based on fine-tuned word embeddings. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, USA.

Todor Mihaylov, Georgi Georgiev, and Preslav Nakov. 2015a. Finding opinion manipulation trolls in news community forums. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, CoNLL '15, pages 310–314, Beijing, China.

Todor Mihaylov, Ivan Koychev, Georgi Georgiev, and Preslav Nakov. 2015b. Exposing paid opinion manipulation trolls. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, RANLP '15, pages 443–450, Hissar, Bulgaria.

Tsvetomila Mihaylova, Pepa Gencheva, Martin Boyanov, Ivana Yovcheva, Todor Mihaylov, Momchil Hardalov, Yasen Kiprov, Daniel Balchev, Ivan Koychev, Preslav Nakov, Ivelina Nikolova, and Galia Angelova. 2016. SUper Team at SemEval-2016 Task 3: Building a feature-rich system for community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, USA.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation*, SemEval '13, pages 321–327, Atlanta, Georgia, USA.

Preslav Nakov, Lluís Màrquez, Walid Magdy, Alessandro Moschitti, Jim Glass, and Bilal Randeree. 2015. SemEval-2015 task 3: Answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 269–281, Denver, Colorado, USA.

Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. SemEval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, USA.

Massimo Nicosia, Simone Filice, Alberto Barrón-Cedeño, Iman Saleh, Hamdy Mubarak, Wei Gao, Preslav Nakov, Giovanni Da San Martino, Alessandro Moschitti, Kareem Darwish, Lluís Màrquez, Shafiq Joty, and Walid Magdy. 2015. QCRI: Answer selection for community question answering - experiments for Arabic and English. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 203–209, Denver, Colorado, USA.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, and Lyle H Ungar. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE*, page e73791.

Aliaksei Severyn and Alessandro Moschitti. 2015. On the automatic learning of sentiment lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '15, pages 1397–1402, Denver, Colorado, USA.

Peter D Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, ACL '02, pages 417–424.